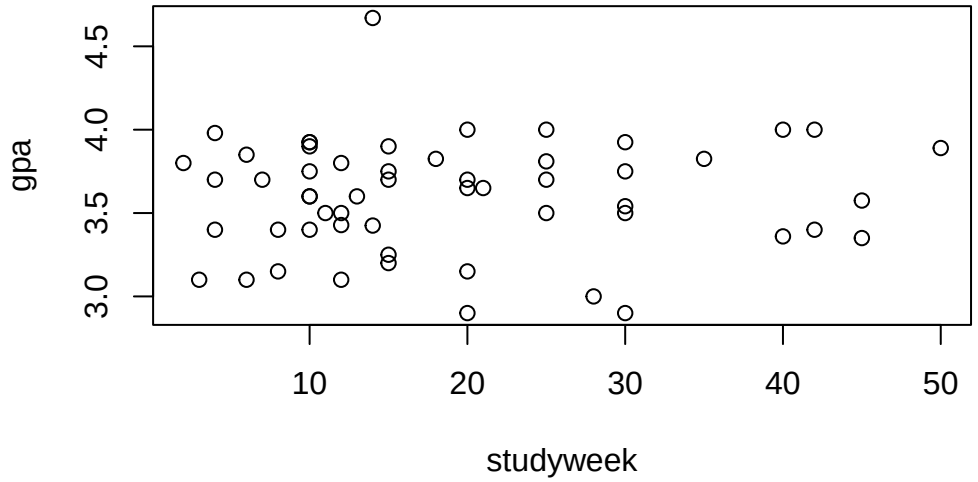
Homework 8

Mark Schulist

1.

```
source("https://www.openintro.org/data/R/gpa.R")
```

```
plot(gpa ~ studyweek, data = gpa)
```




```
cor(gpa$gpa, gpa$studyweek)
```

```
[1] 0.04160403
```

The correlation is very small, basically 0. There is pretty much no relationship between someone's gpa and the amount they study each week.

```
model <- lm(gpa ~ studyweek, data = gpa)
summary(model)
```

```
Call:
lm(formula = gpa ~ studyweek, data = gpa)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71231 -0.18864  0.04784  0.22274  1.07573

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.578490   0.084568  42.315   <2e-16 ***
studyweek   0.001127   0.003719   0.303    0.763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3385 on 53 degrees of freedom
Multiple R-squared:  0.001731,  Adjusted R-squared:  -0.0171
F-statistic: 0.0919 on 1 and 53 DF,  p-value: 0.763
```

The intercept is 3.57, which is the estimated gpa of a student given they have studied zero hours.

The slope is 0.001127, which is the estimated boost to someone's gpa per additional hour of studying.

The r^2 value is 0.001731, which is very small. 0.17% of the variability in someone's gpa can be explained by the number of hours they study per week.

$$H_0 : \text{There is not a linear relationship between gpa and number of hours studied}$$

$$H_a : \text{There is a linear relationship between gpa and number of hours studied}$$

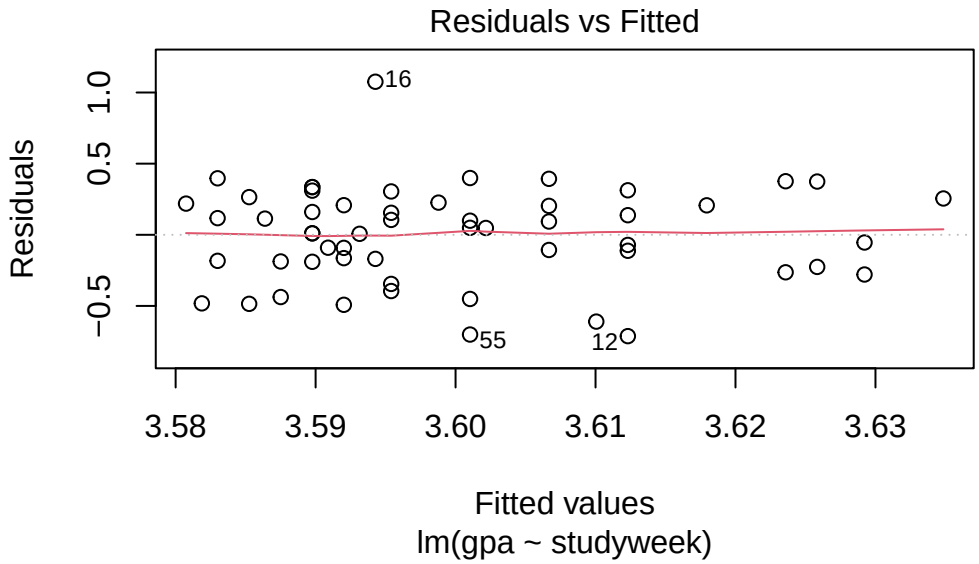
As we saw above, the p-value of the slope is 0.763, which is far above our α. Therefore we do not have evidence to show that there is a linear relationship between gpa and number of hours studied.

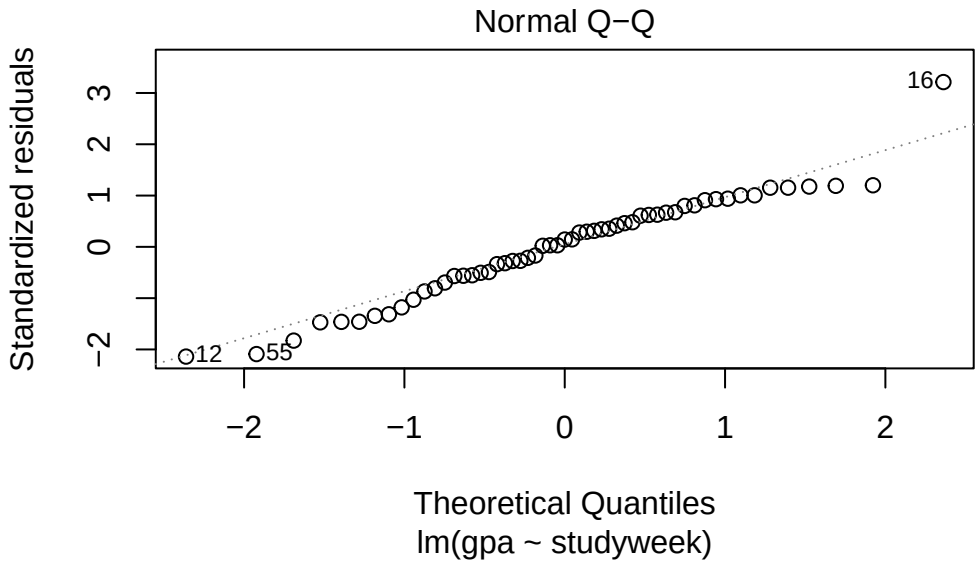
```r
new_studyweek <- data.frame(studyweek = 40)
predict(model, newdata = new_studyweek, interval = "prediction")
```

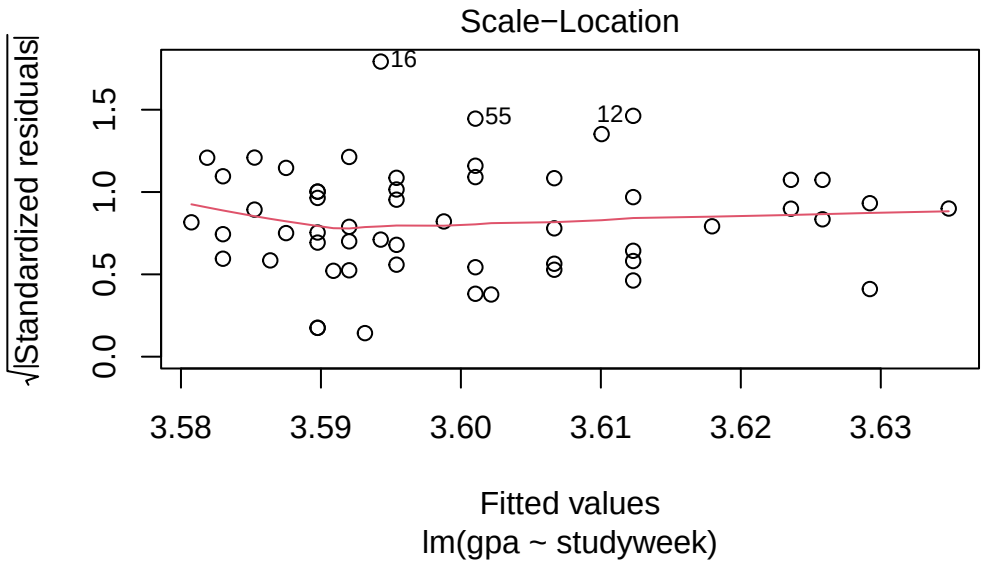
```
       fit      lwr      upr
1 3.623582 2.921103 4.326061
```

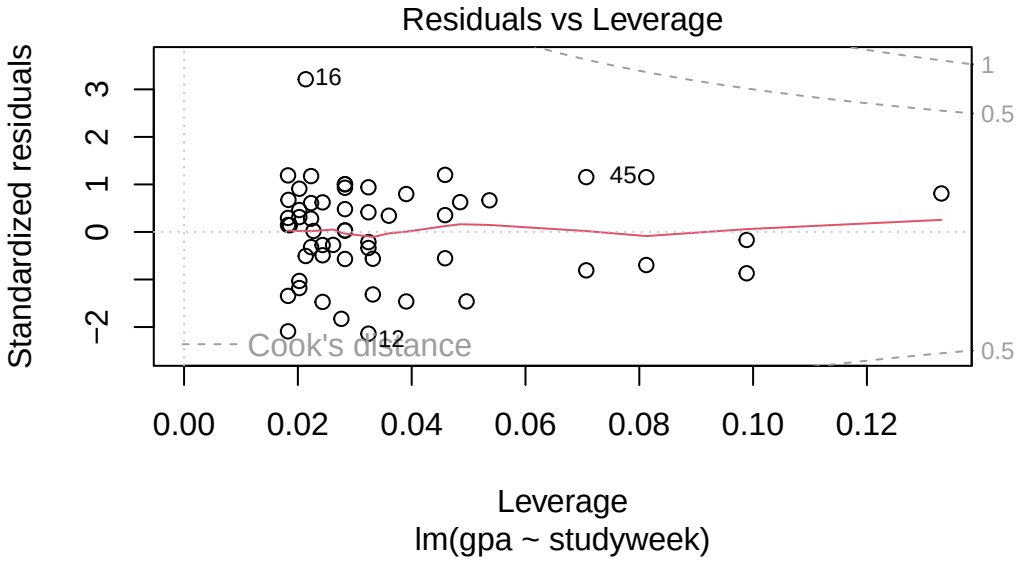
We would predict his GPA to be 3.62. We use a prediction interval because this is a single observation, not the mean of the population. The 95% prediction interval is (2.92, 4.33).

```
plot(model)
```







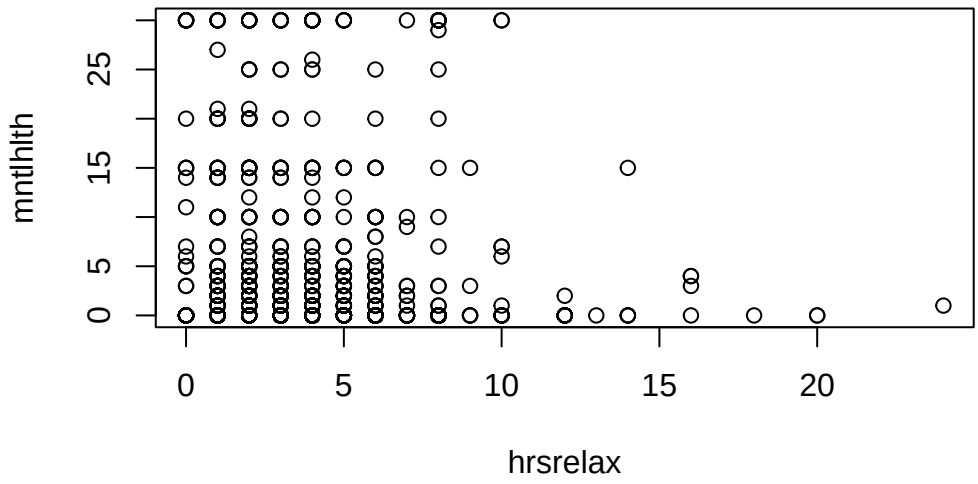


The residuals vs. fitted plot looks like it fits the assumptions of our model. There is even spread around 0. The QQ plot is also very close to the model's assumptions, although near the edges there is a little bit of deviation (although not too bad to say that this model is not worth using).

2.

```
source("https://www.openintro.org/data/R/gss2010.R")
```

```r
plot(mntlhlth ~ hrsrelax, data = gss2010)
```




```r
cor(gss2010$mntlhlth, gss2010$hrsrelax, use = "complete.obs")
```

```
[1] -0.06032906
```

There is a (very) small negative correlation between someone's mental health and the number of hours relaxed.

```r
model1 <- lm(mntlhlth ~ hrsrelax, data = gss2010)
summary(model1)
```

```
Call:
lm(formula = mntlhlth ~ hrsrelax, data = gss2010)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4432 -3.9406 -3.2757  0.3526 27.2321

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.4432     0.3719  11.947   <2e-16 ***
hrsrelax     -0.1675     0.0821  -2.041   0.0415 *
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.287 on 1140 degrees of freedom
  (902 observations deleted due to missingness)
Multiple R-squared:  0.00364,   Adjusted R-squared:  0.002766
F-statistic: 4.164 on 1 and 1140 DF,  p-value: 0.04151
```

The intercept is the estimated number of days an individual would rate their mental health as "not good" when the number of hours they spend on enjoyable activities after work is zero.

The slope is the estimated additional number of days an individual would rate their mental health as "not good" for each additional hour they spend on enjoyable activities after work.

r^2 is 0.00364, which says that 0.364% of the variation in the number of days with bad mental health can be explained by the number of hours spent on enjoyable activities.

H_0: There is not a linear association between the number of days an individual rates their mental health as "not good" and the number of hours they spend on enjoyable activities after work.

H_a: There is a linear association between the number of days an individual rates their mental health as "not good" and the number of hours they spend on enjoyable activities after work.

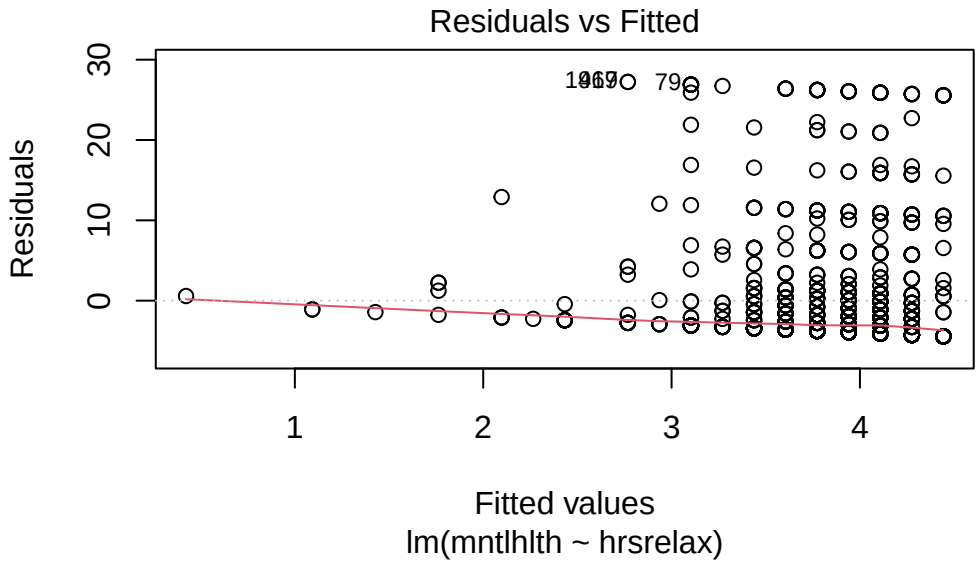
The p-value is 0.0415, so we would (barely) reject the null hypothesis. We have evidence to suggest that there is a linear relationship between the two variables. Although there is statistical significance, the slope is not particularly large. The reason we have significance is because of our large sample size.

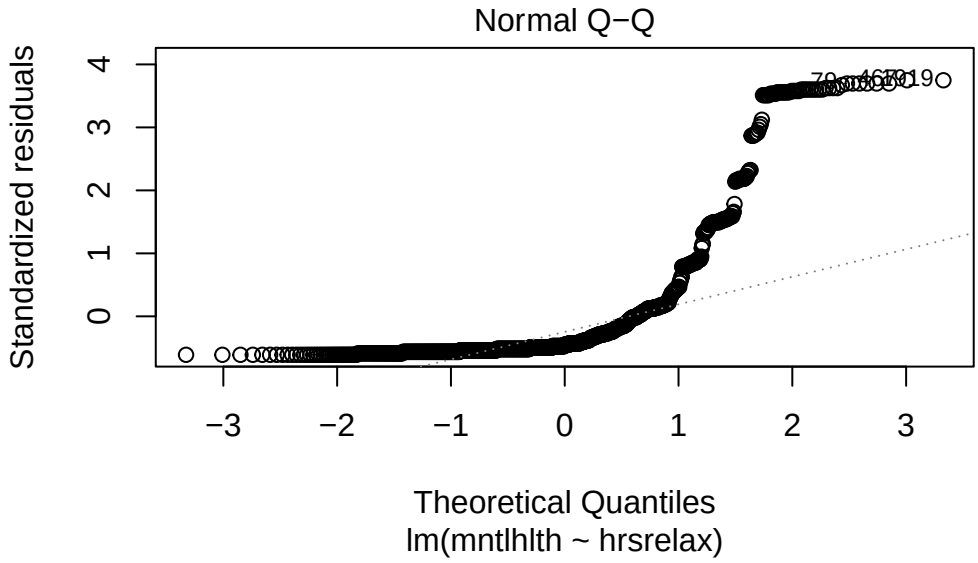
```r
new_hrs <- data.frame(hrsrelax = 5)
predict(model1, newdata = new_hrs, interval = "confidence")
```

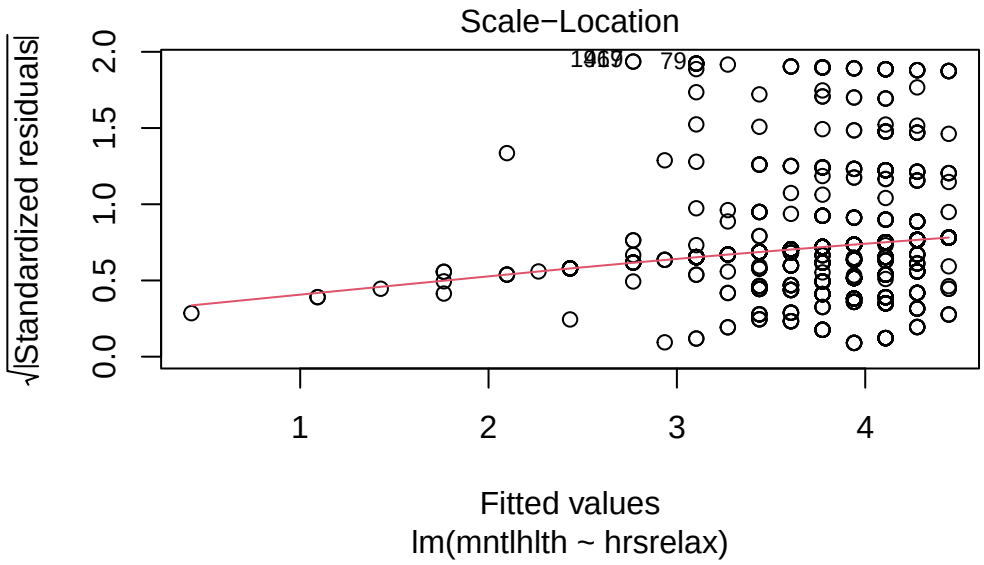
```
       fit      lwr      upr
1 3.605547 3.132834 4.078261
```

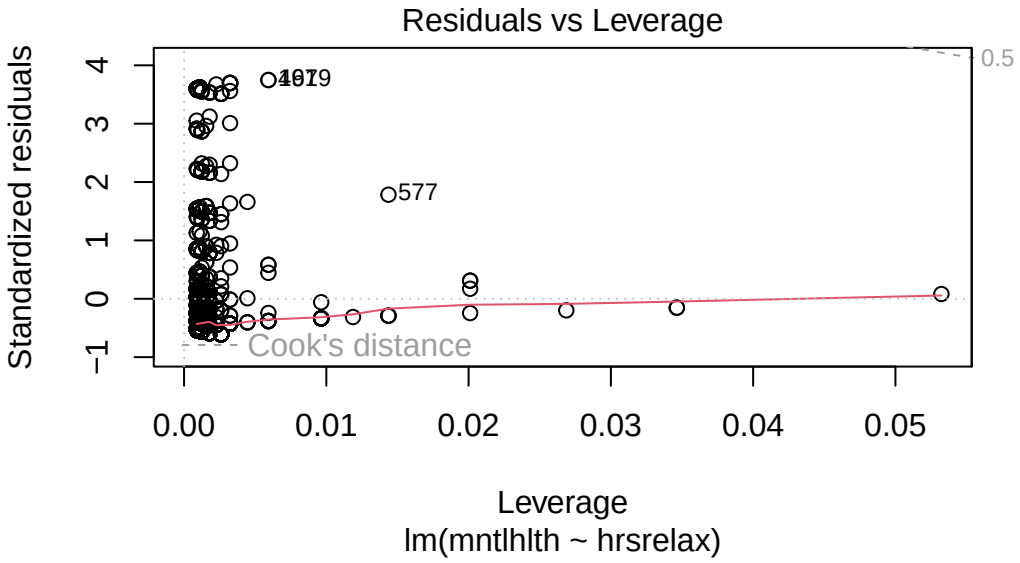
The average number of poor mental health days in a month for individuals who have 5 hours to spend on enjoyable activities is 3.6. The confidence interval is (3.13, 4.08). We use a confidence interval because it is for the population, not a single person.

```r
plot(model1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mntlhlth ~ hrsrelax)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mntlhlth ~ hrsrelax)





These plots reveal that a linear model is definitely not a good model for these data. In the first plot, the residuals are not evenly spread around 0, there are far more on the positive side and for larger fitted values. In the QQ plot, the points are nowhere near the ideal line, showing that the residuals are definitely not normally distributed. These data do not meet the conditions for a single regression analysis.